MODEL SELECTION CRITERION BASED ON THE PREDICTION MEAN SQUARED ERROR IN GENERALIZED ESTIMATING EQUATIONS

Yu Inatsu*1, Shinpei Imori

Department of mathematics, Graduate School of Science, Hiroshima University

ABSTRACT

The present paper considers a model selection criterion in regression models using generalized estimating equation (GEE). Using the prediction mean squared error (PMSE) normalized by the covariance matrix, we propose a new model selection criterion called PMSEG that reflects the correlation between responses. A numerical study reveals that the PMSEG has better performance than previous other criteria for model selection.

Key Words: Generalized estimating equations, Longitudinal data, Prediction mean squared error, Model selection.

1. INTRODUCTION

In real data analysis, correlated data are often discussed in health sciences, medical sciences, economics and many other fields. Longitudinal data, defined from observations of subjects measured repeatedly over time, often arise in these fields as an important example. In general, observations from each subject in longitudinal data are correlated. Liang and Zeger (1986) introduced an extension of the generalized linear model (Nelder and Wedderburn, 1972) to the analysis of longitudinal data, known as the generalized estimating equation (GEE) method. Defining features of the GEE method are that we can use a working (but not necessarily correct) correlation matrix as the correlation matrix, and a full specification of the joint distribution is not required. Hence, the GEE method is widely used in many fields for analyzing longitudinal data.

In addition, the model selection problem in the GEE methodology is also an important. The goodness of fit of the model is commonly measured by a risk function, and the model selection is performed by obtaining a certain estimator of the risk function. For example, the risk function based on the expected Kullback-Leibler (KL) information (Kullback and Leibler, 1951) is often used, and the most famous estimator of the risk function is Akaike’s information criterion (AIC) proposed by Akaike (1973, 1974). The AIC is obtained by using the likelihood, it can be simply

*1 Corresponding author
E-mail address: m126391@hiroshima-u.ac.jp
defined as $\text{AIC} = -2 \times \text{the maximum log likelihood} + 2 \times \text{the number of parameters}$. Furthermore, Nishii (1984), Rao (1988) proposed the GIC as a general extension of the AIC, which is widely applied for selecting the best model, and considered about various properties in many literatures.

However, in the GEE method, the maximum likelihood based model selection criteria such as the AIC or GIC, are not applicable directly because the GEE method is not likelihood based. Some model selection criteria like the AIC or GIC in the GEE method have been already proposed. For example, Pan (2001) proposed the QIC that is based on the quasi likelihood (Wedderburn, 1974). Cantoni et al. (2005) proposed the $GCP$ as a general extension of the Mallows’s $C_p$ (Mallows, 1973). Hin and Wang (2009), Gosho et al. (2011) proposed the CIC to select the correlation structure. Unfortunately, above model selection criteria are not derived by taking account into the correlation between responses. For example, the risk function of the QIC is based on the independent quasi likelihood. In this respect, these criteria are not reflective of the significant feature in longitudinal data.

The principal aim of the present paper is to obtain a new model selection criterion that reflects the correlation between responses. In this study, we have focused on deciding the best subset of variables. By using the risk function based on the prediction mean squared error (PMSE) normalized by the covariance matrix, we propose a new model selection criterion called the PMSEG (the prediction mean squared error in the GEE).

The remainder of the present paper is organized as follows: In Section 2, we consider a stochastic expansion of a GEE estimator. In Section 3, we propose the PMSEG. In Section 4, we verify that the proposed PMSEG has good properties by conducting numerical experiments. In Section 5, we conclude our discussion. Technical details are provided in the Appendix.

2. STOCHASTIC EXPANSION OF THE GEE ESTimator

Let $y_{ij}$ be a scalar response variable, and let $x_{*,ij}$ be a $l$-dimensional nonstochastic vector consists of all possible explanatory variables for the $i$th subject at the $j$th occasion, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Assume that response variables from different subjects are independent and response variables from the same subject are correlated. For $i = 1, \ldots, n$, let $y_i = (y_{i1}, \ldots, y_{im})', X_{*,i} = (x_{*,i1}, \ldots, x_{*,im})'$, and let $X_i = (x_{i1}, \ldots, x_{im})'$ be a $m \times p$ submatrix of the matrix $X_{*,i}$. Liang and Zeger (1986) used the GLM to model the marginal density of $y_{ij}$,

$$f(y_{ij}; x_{ij}, \beta, \phi) = \exp\{y_{ij}\theta_{ij} - a(\theta_{ij})\}/\phi + b(y_{ij}, \phi), \quad (2.1)$$

where, $a(\cdot)$, $b(\cdot)$ are known functions, $\theta_{ij}$ is an unknown location parameter and $\phi$ is a scale parameter. In the GLM framework, the location parameter $\theta_{ij} = u(\eta_{ij}) = \theta_{ij}(\beta)$, where $u(\cdot)$ is a known function, $\eta_{ij} = x_{ij}'\beta$ and $\beta$ is a $p$-dimensional unknown parameter. In the present paper, we assume that $\phi$ is known, and we also assume $\Theta$ to be the natural parameter space (see, Xie and Yang, 2003) of the exponential family of distributions presented in (2.1), and the interior of $\Theta$ is...
denoted as $\Theta^0$. Then $\Theta$ is convex, and in the $\Theta^0$, all derivatives of $a(\cdot)$ and all moments of $y_{ij}$ exist. Under such model conditions, the first two moments of $y_{ij}$ are given by

$$
\mu_{ij}(\beta) = \mathbb{E}[y_{ij}] = \dot{a}(\theta_{ij}), \quad \sigma_{ij}^2(\beta) = \text{Cov}[y_{ij}] = \ddot{a}(\theta_{ij})\phi \equiv \nu(\mu_{ij}(\beta)) \quad \text{(say)}.
$$

In this situation, the expectation of the response $y_{ij}$ is modeled by a link function $g(t) = (\dot{a} \circ u)^{-1}(t)$ and the linear predictor $\eta_{ij}$, i.e., $g(\mu_{ij}(\beta)) = \eta_{ij} = \mathbf{x}'_{ij}\beta$. When $u(s) = s$, we say that $g(t) = (\dot{a})^{-1}(t)$ is the natural link function. For example, the logistic regression model is defined with the natural link function. We call the model with $\mathbf{x}_{*,ij}$ or $\mathbf{x}_{ij}$ as the full or candidate model, respectively. We assume that the true probability density function of $y_{ij}$ can be written as (2.1), i.e., the true model is one of the candidate models.

Denote $\mu_i(\beta) = (\mu_{i1}(\beta), \ldots, \mu_{im}(\beta))^T$, $A_i(\beta) = \text{diag}(\sigma_{i1}^2(\beta), \ldots, \sigma_{im}^2(\beta))$ and $\Delta_i(\beta) = \text{diag}(\partial \theta_{i1}/\partial \eta_{i1}, \ldots, \partial \theta_{im}/\partial \eta_{im})$, where $\text{diag}(a_1, \ldots, a_s)$ denotes the $s \times s$ diagonal matrix whose the $(i, i)$th element is $a_i$. We write $D_i(\beta) = A_i(\beta)\Delta_i(\beta)X_i$, $\Sigma_i(\beta) = A_i^{1/2}(\beta)R^*_iA_i^{1/2}(\beta)$, where $R^*_i$ is the true correlation matrix, and $V_i(\beta, \alpha) = A_i^{1/2}(\beta)R^*_iA_i^{1/2}(\beta)$, where $R_i(\alpha)$ is the working correlation matrix that one can choose freely, which may possibly have a nuisance parameter $\alpha$. Depending on the situation, we can choose some useful working correlation matrices. For example, “Independence” (i.e., $(R)_{jk} = 0$, if $j \neq k$), “Exchangeable” (i.e., $(R)_{jk} = \alpha$, if $j \neq k$), “(first-order) Autoregressive” (i.e., $(R)_{jk} = (R)_{kj} = \alpha^{j-k}$, if $j > k$), “1-dependent” (i.e., $(R)_{jk} = (R)_{kj} = \alpha$, if $j = k + 1$) and “Unstructured” (i.e., $(R)_{jk} = (R)_{kj} = \alpha_{jk}$, if $j > k$). If $R_i(\alpha)$ is equal to $R^*_i$, then $V_i(\beta_0, \alpha) = \Sigma_i(\beta_0) = A_i^{1/2}(\beta_0)R^*_iA_i^{1/2}(\beta_0) = \text{Cov}[y_i]$, where $\beta_0$ is the true value of $\beta$.

Liang and Zeger (1986) proposed the GEE as follows:

$$
s_n(\beta) = \sum_{i=1}^{n} D_i'(\beta)V_i^{-1}(\beta, \alpha)(y_i - \mu_i(\beta)) = 0_p, \quad (2.2)
$$

where $0_p$ is a $p$-dimensional vector of zeros. An estimator $\hat{\beta}$ of $\beta_0$ is defined as a solution of the equation (2.2), and the estimator is called the GEE estimator. In the present paper, we assume that $R_i(\alpha) = R(\alpha)$ and $R^*_i = R_0$. Moreover, to simplify our discussion, we also assume that the nuisance parameter $\alpha$ is known. Hence, we write $V_i(\beta, \alpha) = V_i(\beta)$.

In order to propose a new model selection criterion at Section 3, a stochastic expansion of $\hat{\beta}$ is needed. In this section, we obtain the stochastic expansion of $\hat{\beta}$ up to the order $n^{-1}$. For simplicity, we omit $(\beta)$ from functions of $\beta$ like $\mu_{ij}(\beta) = \mu_{ij}$. Furthermore, in order to distinguish a function of $\beta$ evaluated at the true parameter $\beta_0$ and GEE estimator $\hat{\beta}$, we write such as $\mu_{ij}(\beta_0) = \mu_{ij,0}$ and $\mu_{ij}(\hat{\beta}) = \hat{\mu}_{ij}$, respectively. Furthermore, in order to assure asymptotic properties of the GEE estimator, we consider the following regularity assumptions (see, e.g., Xie and Yang, 2003):

C1. $\beta_0$ is in an admissible set $B$, where $B$ is an open set in $\mathbb{R}^p$ for the parameter $\beta$.

C2. $\mathbf{x}'_{ij}\beta \in g(\mathcal{M})$ for all $\beta \in B$, where $\mathcal{M}$ is the image of $\dot{a}(\Theta^0)$. 

3
C3. $u(\eta_{ij})$ is four times continuously differentiable and $u(\eta_{ij}) > 0$ in $g(M)^0$.

C4. $H_{n,0}$ and $M_{n,0}$ are positive definite when $n$ is large, where

$$H_n = \sum_{i=1}^{n} D_i' V_i^{-1} D_i, \quad M_n = \sum_{i=1}^{n} D_i' V_i^{-1} \Sigma_i V_i^{-1} D_i.$$  

Condition C1 and C2 are necessary to have the GLM for all $\beta$. Condition C3 and C4 are necessary to calculate the bias. In addition, in order to assure the strong consistency, asymptotic normality and uniqueness of the GEE estimator, we consider the following additional assumptions, which can be derived slightly modifying the results reported by Xie and Yang (2003):

C5. $\liminf_{n \to \infty} \lambda_{\min}(H_{n,0}/n) > 0$, where $\lambda_{\min}(A)$ is the smallest eigenvalue of symmetric matrix $A$.

C6. Sequence $\{x_{ij}\}$ lies in $\mathcal{X}$ with $u(x_{ij}\beta) \in \Theta^0$, $\beta \in \mathcal{B}$, where $\mathcal{X}$ is a compact set for regressors $x_{ij}$.

C7. In a neighborhood of $\beta_0$, say $N$, there exist a constant $c_0 > 0$ and some $\delta > 0$, independent of $n$, such that, when $n \to \infty$, for any $p$-dimensional vector $\lambda$, $||\lambda|| = 1$,

$$\lambda' \frac{\partial s_n}{\partial \beta'} \lambda \geq c_0 \lambda_{\max}^{(1/2)+\delta}(M_{n,0}), \quad \text{a.s. for } \beta \in N.$$  

C8. The equation (2.2) has a unique solution when $n$ is large.

Here $\lambda_{\max}(A)$ is the largest eigenvalue of symmetric matrix $A$. According to THEOREM 7 and COROLLARY 1 in Xie and Yang (2003), $\hat{\beta}$ has the strong consistency and asymptotic normality under these conditions. Furthermore, from C5, $H_{n,0} = O(n)$.

Based on the above conditions, to perform the stochastic expansion of $\hat{\beta}$, we focus on the fact that $\hat{s}_n = 0_p$. By applying a Taylor expansion around $\hat{\beta} = \beta_0$ to this equation, the GEE is expanded as follows:

$$0_p = s_{n,0} + \frac{\partial s_n}{\partial \beta'} \bigg|_{\beta = \beta_0} (\hat{\beta} - \beta_0) + \frac{1}{2} ((\hat{\beta} - \beta_0)' \otimes I_p) \left( \frac{\partial}{\partial \beta} \otimes \frac{\partial s_n}{\partial \beta'} \right) \bigg|_{\beta = \beta_0} (\hat{\beta} - \beta_0)$$

$$= s_{n,0} - H_{n,0}(I_p + G_{1,0} + G_{2,0})(\hat{\beta} - \beta_0) + \frac{1}{2} ((\hat{\beta} - \beta_0)' \otimes I_p) L_1(\beta^*)(\hat{\beta} - \beta_0),$$  

(2.3)

where $\beta^*$ lies between $\beta_0$ and $\hat{\beta}$, $I_p$ is a $p$-dimensional identity matrix, $G_{1,0}, G_{2,0}$ and $L_1(\beta^*)$ are defined by

$$G_{1,0} = -H_{n,0}^{-1} \sum_{i=1}^{n} D_i' \left( \frac{\partial}{\partial \beta'} \otimes V_i^{-1} \bigg|_{\beta = \beta_0} \right) \{I_p \otimes \{y_{i} - \mu_{i,0}\}\},$$  

$$G_{2,0} = -H_{n,0}^{-1} \sum_{i=1}^{n} \left( \frac{\partial}{\partial \beta'} \otimes D_i' \bigg|_{\beta = \beta_0} \right) \{I_p \otimes \{V_i^{-1}(y_{i} - \mu_{i,0})\}\},$$  

$$L_1(\beta^*) = \left( \frac{\partial}{\partial \beta} \otimes \frac{\partial s_n}{\partial \beta'} \right) \bigg|_{\beta = \beta^*}.$$
Note that for a matrix $W = (w_{ij})$, the derivative of $W$ by $\beta = (\beta_1, \ldots, \beta_p)'$ and $\beta_k$ are respectively defined as follows:

$$\frac{\partial}{\partial \beta'} \otimes W = \left( \frac{\partial W}{\partial \beta_1}, \ldots, \frac{\partial W}{\partial \beta_p} \right), \quad \frac{\partial}{\partial \beta} \otimes W = \left( \frac{\partial W'}{\partial \beta_1}, \ldots, \frac{\partial W'}{\partial \beta_p} \right), \quad \frac{\partial W}{\partial \beta_k} = \left( \frac{\partial w_{ij}}{\partial \beta_k} \right).$$

Also note that $L_1(\beta^*) = O_p(n)$, $s_{n,0} = O_p(n^{1/2})$, $\hat{\beta} - \beta_0, G_{1,0}, G_{2,0} = O_p(n^{-1/2})$. Thus, (2.3) yields

$$\hat{\beta} - \beta_0 = H_{n,0}^{-1} s_{n,0} + O_p(n^{-1}) = b_{1,0} + O_p(n^{-1}). \quad (2.4)$$

Similarly, the GEE can also be expanded as follows:

$$s_{n,0} = H_{n,0}(I_p + G_{1,0} + G_{2,0})(\hat{\beta} - \beta_0) - \frac{1}{2} \{ (\hat{\beta} - \beta_0)' \otimes I_p \} \{ G_{3,0} + (L_1(\beta_0) - G_{3,0}) \}(\hat{\beta} - \beta_0)$$

$$- \frac{1}{6} \{ (\hat{\beta} - \beta_0)' \otimes I_p \} \left\{ \frac{\partial}{\partial \beta'} \otimes \left( \frac{\partial}{\partial \beta} \otimes \frac{\partial s_n}{\partial \beta} \right) \right\}_{\beta = \hat{\beta}^*} \{ (\hat{\beta} - \beta_0) \otimes (\hat{\beta} - \beta_0) \}, \quad (2.5)$$

where $\beta^*$ lies between $\beta_0$ and $\hat{\beta}$, and $G_{3,0} = E[L_1(\beta_0)].$

Note that the order of the last term of equation (2.5) is $O_p(n^{-1/2})$. Also note that $G_{3,0} = O(n)$ and $L_1(\beta_0) - G_{3,0} = O_p(n^{1/2})$. Therefore, by using equation (2.4) and (2.5), $\hat{\beta}$ can be expanded as follows:

$$\hat{\beta} - \beta_0 = b_{1,0} + b_{2,0} + O_p(n^{-3/2}), \quad (2.6)$$

where $b_{2,0} = H_{n,0}^{-1}(b_{1,0}' \otimes I_p)G_{3,0}b_{1,0}/2 - G_{1,0}b_{1,0} - G_{2,0}b_{1,0}$. Note that $b_{1,0}$ and $b_{2,0}$ are $O_p(n^{-1/2})$ and $O_p(n^{-1})$, respectively.

### 3. MAIN RESULT

In this section, we propose a new model selection criterion. The goodness of fit of the model is measured by the risk function based on the PMSE normalized by the covariance matrix as follows:

$$\text{Risk}_p = \text{PMSE} - mn = E_y \left[ E_z \left( \sum_{i=1}^n (z_i - \hat{\mu}_i)' \Sigma_{i,0}^{-1} (z_i - \hat{\mu}_i) \right) \right] - mn,$$

where $z_i = (z_{i1}, \ldots, z_{im})'$ is an $m$-dimensional random vector that is independent of $y_i$ and has the same distribution as $y_i$. It is easy to show that if $\hat{\beta}$ is $\beta_0$, then Risk$_p$ has the minimum value of zero, i.e., the PMSE has the minimum value of $mn$. For this reason, we would like to select the model, which has the minimum PMSE. However, since the PMSE is typically unknown, we must estimate it.
We define \( R_0(\beta) , \mathcal{L}(\beta_1, \beta_2) \) and \( \mathcal{L}^*(\beta) \) as follows:

\[
R_0(\beta) = \frac{1}{n} \sum_{i=1}^{n} A_i^{-1/2}(y_i - \mu_i)(y_i - \mu_i)'A_i^{-1/2},
\]

\[
\mathcal{L}(\beta_1, \beta_2) = \sum_{i=1}^{n}(y_i - \mu_i(\beta_1))'A_i^{-1/2}(\beta_2)R_0^{-1}(\beta_2)A_i^{-1/2}(\beta_2)(y_i - \mu_i(\beta_1)),
\]

\[
\mathcal{L}^*(\beta) = \sum_{i=1}^{n}(y_i - \mu_i)'\Sigma_{i,0}^{-1}(y_i - \mu_i).
\]

Based on the above, we estimate PMSE by \( \mathcal{L}(\hat{\beta}, \hat{\beta}_f) \), where \( \hat{\beta}_f \) is a GEE estimator that is obtained under the “full” model. Specifically, \( \hat{\beta}_f \) is defined as the solution to the following equation:

\[
s_{f,n}(\beta_*) = \sum_{i=1}^{n} D_i'(\beta_*)V_i^{-1}(\beta_*, \alpha^*)(y_i - \mu_*(\beta_*)) = 0_i,
\]

where \( D_i(\beta_*) = A_i(\beta_*)\Delta_i(\beta_*)X_{i,i} \), \( V_i(\beta_*, \alpha^*) = A_i^{1/2}(\beta_*)^\prime \tilde{R}_i(\alpha^*)A_i^{1/2}(\beta_*) \) and \( \tilde{R}_i(\alpha^*) \) is a positive definite working correlation matrix that one can choose freely. We assume that the nuisance parameter \( \alpha^* \) is known. Note that \( \beta_* \) is an \( l \)-dimensional unknown parameter under the full model. Also note that \( \tilde{R}_i(\alpha^*) \) is the same for all candidate models. The reason for using \( \hat{\beta}_f \) is discussed later. For simplicity, we write \( \mathcal{L}(\beta_0, \beta_2) = \mathcal{L}(\beta_2) \) and \( \mathcal{L}^*(\beta_0) = \mathcal{L}^* \).

Since \( \mathcal{L}(\hat{\beta}, \hat{\beta}_f) \) is not an asymptotically unbiased estimator of PMSE, we evaluate the asymptotic bias in order to propose the new model selection criterion. The bias when we estimate the PMSE by \( \mathcal{L}(\hat{\beta}, \hat{\beta}_f) \) is given as

\[
\text{Bias} = \text{PMSE} - E_y[\mathcal{L}(\hat{\beta}, \hat{\beta}_f)] = \{\text{Risk}_p - E_y[\mathcal{L}^*(\hat{\beta})]\} + \{E_y[\mathcal{L}^*(\hat{\beta}) - \mathcal{L}^*]\} + \{E_y[\mathcal{L}^*(\hat{\beta}_f)] - E_y[\mathcal{L}^*(\hat{\beta}_f) - \mathcal{L}^*]\}
\]

\[
= \text{Bias1} + \text{Bias2} + \text{Bias3} + \text{Bias4}.
\]

(3.1)

Here, we can see that Bias3 in (3.1) satisfies

\[
\text{Bias3} = E_y \left[ \sum_{i=1}^{n} (y_i - \mu_{i,0})' \left( \Sigma_{i,0}^{-1} - A_i^{-1/2}(\hat{\beta}_f)R_0^{-1}(\hat{\beta}_f)A_i^{-1/2}(\hat{\beta}_f) \right) (y_i - \mu_{i,0}) \right]
\]

\[
= mn - E_y \left[ \sum_{i=1}^{n} (y_i - \mu_{i,0})' A_i^{-1/2}(\hat{\beta}_f)R_0^{-1}(\hat{\beta}_f)A_i^{-1/2}(\hat{\beta}_f)(y_i - \mu_{i,0}) \right].
\]

Therefore, we can ignore the calculation of Bias3 because it is not dependent on candidate models.

Similarly, Bias1 in (3.1) is expanded as

\[
\text{Bias1} = E_y \left[ E_z \left[ \sum_{i=1}^{n} (z_i - \bar{\mu}_i)' \Sigma_{i,0}^{-1}(z_i - \bar{\mu}_i) \right] - \sum_{i=1}^{n} (y_i - \bar{\mu}_i)' \Sigma_{i,0}^{-1}(y_i - \bar{\mu}_i) \right]
\]

\[
= 2E_y \left[ \sum_{i=1}^{n} (y_i - \mu_{i,0})' \Sigma_{i,0}^{-1}(\mu_i - \mu_{i,0}) \right].
\]

(3.2)
Applying a Taylor expansion around \( \hat{\beta} = \beta_0 \) to \( \hat{\mu}_i \) yields

\[
\hat{\mu}_i = \mu_{i,0} + \left. \frac{\partial \mu_i}{\partial \beta} \right|_{\beta=\beta_0} (\hat{\beta} - \beta_0) + \frac{1}{2} \left\{ (\hat{\beta} - \beta_0)' \otimes I_m \right\} \left. \left( \frac{\partial}{\partial \beta} \otimes \frac{\partial \mu_i}{\partial \beta} \right) \right|_{\beta=\beta_0} (\hat{\beta} - \beta_0)
\]

\[
+ \frac{1}{6} \left\{ (\hat{\beta} - \beta_0)' \otimes I_m \right\} \left. \left( \frac{\partial}{\partial \beta} \otimes \frac{\partial}{\partial \beta} \frac{\partial \mu_i}{\partial \beta} \right) \right|_{\beta=\beta_0} \left\{ (\hat{\beta} - \beta_0) \otimes (\hat{\beta} - \beta_0) \right\}
\]

\[
= \mu_{i,0} + D_{i,0}(\hat{\beta} - \beta_0) + \frac{1}{2} \left\{ (\hat{\beta} - \beta_0)' \otimes I_m \right\} D_{i,0}^{(1)}(\hat{\beta} - \beta_0) + O_p(n^{-3/2}), \quad (3.3)
\]

Here, \( \beta^{***} \) lies between \( \beta_0 \) and \( \hat{\beta} \). Substituting the stochastic expansion of \( \hat{\beta} \) in (2.6) into (3.3) yields the following:

\[
\hat{\mu}_i - \mu_{i,0} = D_{i,0}b_{i,0} + \left\{ D_{i,0}b_{2,0} + \frac{1}{2}(b'_{1,0} \otimes I_m)D_{i,0}^{(1)}b_{1,0} \right\} + O_p(n^{-3/2}). \quad (3.4)
\]

By combining (3.2) and (3.4), we obtain

\[
\frac{1}{2} \text{Bias}^1 = E_y \left[ \sum_{i=1}^{n} (y_i - \mu_{i,0})' \Sigma^{-1}_{i,0} D_{i,0}b_{1,0} \right]
\]

\[
+ E_y \left[ \sum_{i=1}^{n} (y_i - \mu_{i,0})' \Sigma^{-1}_{i,0} \left\{ D_{i,0}b_{2,0} + \frac{1}{2}(b'_{1,0} \otimes I_m)D_{i,0}^{(1)}b_{1,0} \right\} \right]
\]

\[
+ E_y[O_p(n^{-1/2})]. \quad (3.5)
\]

Note that \( E[(y_i - \mu_{i,0})' (y_j - \mu_{j,0})] = 0, \ (i \neq j) \), the first term of (3.5) can be calculated as

\[
E_y \left[ \sum_{i=1}^{n} (y_i - \mu_{i,0})' \Sigma^{-1}_{i,0} D_{i,0}b_{1,0} \right] = E_y \left[ \sum_{i=1}^{n} \sum_{j=1}^{n} (y_i - \mu_{i,0})' \Sigma^{-1}_{i,0} D_{i,0} H_{n,0}^{-1} \dot{D}_{j,0} V_{j,0}^{-1} (y_j - \mu_{j,0}) \right]
\]

\[
= E_y \left[ \sum_{i=1}^{n} (y_i - \mu_{i,0})' \Sigma^{-1}_{i,0} D_{i,0} H_{n,0}^{-1} \dot{D}_{i,0} V_{i,0}^{-1} (y_i - \mu_{i,0}) \right]
\]

\[
= \text{tr} \left[ H_{n,0}^{-1} \sum_{i=1}^{n} D_{i,0} V_{i,0}^{-1} D_{i,0} \right] = \text{tr}[H_{n,0}^{-1} H_{n,0}] = p. \quad (3.6)
\]

Similarly, since \( E[(y_i - \mu_{i,0}) \otimes (y_j - \mu_{j,0})'(y_k - \mu_{k,0})] = 0_{n^2} \), (not \( i = j = k \)), the second term of (3.5) can be expanded as

\[
E_y \left[ \sum_{i=1}^{n} (y_i - \mu_{i,0})' \Sigma^{-1}_{i,0} \left\{ D_{i,0}b_{2,0} + \frac{1}{2}(b'_{1,0} \otimes I_m)D_{i,0}^{(1)}b_{1,0} \right\} \right]
\]

\[
= E_y \left[ \sum_{i=1}^{n} (y_i - \mu_{i,0})' \Sigma^{-1}_{i,0} \left\{ D_{i,0}b_{2,0} + \frac{1}{2}(b'_{1,0} \otimes I_m)D_{i,0}^{(1)}b_{1,0} \right\} ,
\]

\[
\text{tr}[H_{n,0}^{-1} H_{n,0}] = p.
\]
where \( b_{1i,0} = H_{n,0}^{-1}D'_{i,0}V_{i,0}^{-1}(y_i - \mu_i,0) \), \( b_{2i,0} = H_{n,0}^{-1}(b'_{1i,0} \otimes I_p)G_{3,0}b_{1i,0}/2 - G_{1i,0}b_{1i,0} - G_{2i,0}b_{1i,0} \),

\[
G_{1i,0} = -H_{n,0}^{-1}D'_{i,0}\left(\frac{\partial}{\partial \beta'} \otimes V_i^{-1}\right)\left|_{\beta = \beta_0}\right\{I_p \otimes (y_i - \mu_i,0)\},
\]

\[
G_{2i,0} = -H_{n,0}^{-1}\left(\frac{\partial}{\partial \beta'} \otimes D_i\right)\left|_{\beta = \beta_0}\right\} \{I_p \otimes \{V_i^{-1}(y_i - \mu_i,0)\}\}.
\]

Note that \( D_{i,0}b_{2i,0} + (b'_{1i,0} \otimes I_m)D_{i,0}^{(1)}b_{1i,0}/2 = O_p(n^{-2}) \), the second term of (3.5) can be obtained as

\[
\text{E}_y \left[ \sum_{i=1}^{n} (y_i - \mu_i,0)'\Sigma^{-1}_{i,0}\left\{D_{i,0}b_{2,0} + \frac{1}{2}(b'_{1,0} \otimes I_m)D_{i,0}^{(1)}b_{1,0}\right\} \right] = O(n^{-1}). \tag{3.7}
\]

Under certain conditions, the limit of the expectation is equal to the expectation of the limit. Furthermore, in many cases of practical interest, a moment of statistic can be expanded as power series in \( n^{-1} \) (see e.g., Hall, 1992). Therefore, by substituting (3.6) and (3.7) into (3.5), we obtain the asymptotic expansion of Bias1 up to order 1 as

\[
\text{Bias1} = 2p + O(n^{-1}). \tag{3.8}
\]

Moreover, by using the same argument of the calculation of Bias1, we obtain

\[
\text{Bias2} + \text{Bias4} = O(n^{-1}). \tag{3.9}
\]

The derivation of (3.9) is shown in Appendix.

Consequently, by substituting (3.8) and (3.9) into (3.1), we obtain the asymptotic expansion of Bias up to order 1 as

\[
\text{Bias} = 2p + \text{Bias3} + O(n^{-1}). \tag{3.10}
\]

Recall that Bias3 is not dependent on candidate models. Hence, the PMSEG can then be defined by adding an estimated (Bias – Bias3) to \( \mathcal{L}(\hat{\beta}, \hat{\beta}_f) \), i.e.,

\[
\text{PMSEG} = \mathcal{L}(\hat{\beta}, \hat{\beta}_f) + 2p. \tag{3.11}
\]

(3.11) is our proposed model selection criterion called the PMSEG (the prediction mean squared error in the GEE). Recall that \( \hat{\beta}_f \) is estimated under the full model and it is not dependent on candidate models. Since the covariance matrix in the PMSEG is estimated by \( \hat{\beta}_f \), the PMSEG can be simply defined. If the covariance matrix is estimated by \( \hat{\beta} \), Bias3 in (3.11) is different for each candidate model. Unfortunately, it is difficult and too expensive to calculate Bias3. This is one of the advantages of using \( \hat{\beta}_f \) for estimating the covariance matrix. For actual use, we recommend to use the working independence matrix in order to obtain \( \hat{\beta}_f \) since we can get \( \hat{\beta}_f \) easily by omitting the calculations of the working correlation matrix. Fitzmaurice (1995) mentioned that the GEE estimator under the working independence assumption is often inefficient. Nevertheless, from some simulation results, we confirmed that the estimation of the covariance matrix using the inefficient estimator does not dramatically influence the performance of the PMSEG.
4. NUMERICAL STUDIES

In this section, we confirm a usefulness of the PMSEG through comparisons with the QIC, which is a representative previous study. The QIC is an estimator of a risk function based on the quasi likelihood under the independence assumption. The risk function that is estimated by the QIC, which is called Risk_Q in this paper, and the quasi likelihood Q(\cdot) are defined as follows:

\[
\text{Risk}_Q = E_y \left[ E_z \left[ -2 \sum_{i=1}^{n} \sum_{j=1}^{m} Q(\hat{\beta}; z_{ij}) \right] \right], \quad Q(\hat{\beta}; z_{ij}) = \int_{z_{ij}}^{g^{-1}(x'_i\beta)} \frac{z_{ij} - t}{\nu(t)} dt,
\]

where \(z_i = (z_{i1}, \ldots, z_{im})'\) is an \(m\)-dimensional random vector that is independent of \(y_i\) and has the same distribution as \(y_i\). Note that since the Risk_Q is considered under the independence assumption, both the Risk_Q and QIC are not reflective of the correlation between responses. Also note that the PMSEG and QIC are estimators of the Risk_P and Risk_Q, respectively. In general, the Risk_P and Risk_Q are different, i.e., the PMSEG and QIC are criteria from different viewpoints.

At the beginning, we examine the numerical studies for the frequencies of candidate models and the prediction error of the best models selected by the PMSEG and QIC. We prepared the fifteen candidate models with \(n = 100\) and \(m = 3\). First, we constructed a \(3 \times 5\) explanatory variable matrix \(X_{s,i} = (x_{s,i1}, x_{s,i2}, x_{s,i3})'\), \(i = 1, \ldots, 100\). The first column of \(X_{s,i}\) is \(1_3\), where \(1_3\) is a 3-dimensional vector of ones, and the second column of \(X_{s,i}\) is \(1_3 \times \varsigma_i\), where \(\varsigma_1, \ldots, \varsigma_{100}\) are i.i.d. binomial distribution \(B(1, 0.5)\). The third column of \(X_{s,i}\) is \((0, 1, 2)'\), and the remaining six elements of \(X_{s,i}\) were defined by realizations of independent variables with uniform distribution on the interval \([-1, 1]\).

In this simulation, we prepared two situations, as follows:

Case 1: \(\text{Corr}[y_{ij}, y_{ij'}] = 0.85|j-j'|, \quad \beta_0 = (0.25, -0.25, -0.25, 0, 0)'\),

Case 2: \(\text{Corr}[y_{ij}, y_{ij'}] = 0.35, \quad \beta_0 = (0.25, -0.25, -0.25, 0, 0)'\).

The explanatory variables matrix for the \(i\)th subject in the \((5k + k^*)\)th model consists of the first \(k^*\) column of \(X_{s,i}, k = 0, 1, 2, \ k^* = 1, \ldots, 5\). For the working correlation matrix, we prepare three different matrices, working exchangeable matrix \((k = 0)\), working autoregressive matrix \((k = 1)\) and working independence matrix \((k = 2)\). Thus, in case 1, the true model is the eighth model, and in case 2, the true model is the third model. We simulated 10,000 realizations of \(y = (y_{11}, \ldots, y_{1m}, \ldots, y_{100,1}, \ldots, y_{100,m})'\) in the logistic regression model, i.e., \(y_{ij} \sim \text{indep. } B(1, p_{ij})\), where \(p_{ij} = \text{logit}^{-1}(x'_{ij}\beta_0), \ i = 1, \ldots, 100, \ j = 1, 2, 3\). Note that we used the working independence matrix for obtaining \(\hat{\beta}_f\) in this simulation. Tables 1 and 2 list the following characteristics.

Table 1: Selection probability and prediction error for the Case 1

<table>
<thead>
<tr>
<th>Criterion</th>
<th>$k^*$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>SP_C</th>
<th>SP_M</th>
<th>PEB_P</th>
<th>PEB_Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMSEG</td>
<td>Exchangeable</td>
<td>0.95</td>
<td>0.08</td>
<td>11.02</td>
<td>2.62</td>
<td>1.10</td>
<td></td>
<td></td>
<td>70.49</td>
<td>4.37</td>
</tr>
<tr>
<td></td>
<td>Autoregressive</td>
<td>2.61</td>
<td>0.48</td>
<td>50.71</td>
<td>9.77</td>
<td>6.92</td>
<td></td>
<td></td>
<td>95.61</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Independence</td>
<td>0.22</td>
<td>0.05</td>
<td>12.72</td>
<td>0.67</td>
<td>0.08</td>
<td></td>
<td></td>
<td>5.04</td>
<td>10.54</td>
</tr>
<tr>
<td>QIC</td>
<td>Exchangeable</td>
<td>4.36</td>
<td>0.01</td>
<td>2.60</td>
<td>3.46</td>
<td>3.60</td>
<td></td>
<td></td>
<td>72.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Autoregressive</td>
<td>38.36</td>
<td>0.26</td>
<td>22.20</td>
<td>6.11</td>
<td>5.86</td>
<td></td>
<td></td>
<td>55.50</td>
<td>10.54</td>
</tr>
<tr>
<td></td>
<td>Independence</td>
<td>1.51</td>
<td>0.00</td>
<td>3.09</td>
<td>5.17</td>
<td>3.41</td>
<td></td>
<td></td>
<td>41.89</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Selection probability and prediction error for the Case 2

<table>
<thead>
<tr>
<th>Criterion</th>
<th>$k^*$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>SP_C</th>
<th>SP_M</th>
<th>PEB_P</th>
<th>PEB_Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMSEG</td>
<td>Exchangeable</td>
<td>13.97</td>
<td>0.89</td>
<td>26.28</td>
<td>8.06</td>
<td>6.76</td>
<td></td>
<td></td>
<td>55.96</td>
<td>4.88</td>
</tr>
<tr>
<td></td>
<td>Autoregressive</td>
<td>7.67</td>
<td>1.77</td>
<td>9.10</td>
<td>1.97</td>
<td>1.20</td>
<td></td>
<td></td>
<td>72.79</td>
<td>55.50</td>
</tr>
<tr>
<td></td>
<td>Independence</td>
<td>2.06</td>
<td>0.85</td>
<td>17.00</td>
<td>1.73</td>
<td>0.69</td>
<td></td>
<td></td>
<td>10.54</td>
<td>5.53</td>
</tr>
<tr>
<td>QIC</td>
<td>Exchangeable</td>
<td>29.70</td>
<td>1.14</td>
<td>20.06</td>
<td>7.46</td>
<td>5.64</td>
<td></td>
<td></td>
<td>64.00</td>
<td>57.93</td>
</tr>
<tr>
<td></td>
<td>Autoregressive</td>
<td>2.55</td>
<td>0.19</td>
<td>1.70</td>
<td>1.17</td>
<td>1.14</td>
<td></td>
<td></td>
<td>55.96</td>
<td>55.96</td>
</tr>
<tr>
<td></td>
<td>Independence</td>
<td>7.53</td>
<td>0.96</td>
<td>16.64</td>
<td>2.56</td>
<td>1.56</td>
<td></td>
<td></td>
<td>10.54</td>
<td>5.53</td>
</tr>
</tbody>
</table>

of the model selected by the criterion as the best model, which are respectively estimated as

$$PEB_P = \frac{1}{10000} \sum_{v=1}^{10000} E_z \left[ \sum_{i=1}^{n} (z_i - \mu_i(\hat{\beta}_{B_v}))' \Sigma_{i,0}^{-1}(z_i - \mu_i(\hat{\beta}_{B_v})) \right] - mn,$$

$$PEB_Q = \frac{1}{10000} \sum_{v=1}^{10000} E_z \left[ -2 \sum_{i=1}^{n} \sum_{j=1}^{m} Q(\hat{\beta}_{B_v}; z_{ij}) \right].$$

(2) Selection probability: the frequency of the best model chosen by minimizing each criterion. In particular, the SP_C/SP_M is the frequency that the working correlation matrix/mean structure of the selected model is correctly specified.

Here $z_i$ is a future observation, and $\hat{\beta}_{B_v}$ is the value of $\hat{\beta}$ of the selected model at the $v$th iteration. In particular, both the PEB_P and PEB_Q are important properties because these are equivalent to the Risk_P and Risk_Q of the best model selected by the criterion, respectively. We would like to note that the PMSEG selects the model which minimizes the Risk_P, and the QIC selects the model which minimizes the Risk_Q. Thus, the model selected by the PMSEG does not necessarily minimize the Risk_Q and the model selected by the QIC does not necessarily minimize the Risk_P. In other words, in order to evaluate the goodness of the criterion, the PEB_P and PEB_Q are favorable indicators for the PMSEG and QIC, respectively.

From Tables 1 and 2, we can see that the value of the PEB_P from the model selected by the
PMSEG is smaller than that from the model selected by the QIC. This result is justified since the PEB_P is the favorable indicator for the PMSEG. However, surprisingly, although the PEB_Q is the favorable indicator for the QIC, the value of the PEB_Q from the model selected by the PMSEG is also smaller. This result means that the PMSEG is better than the QIC whether evaluating the goodness of the criterion by the PEB_P or PEB_Q. Moreover, both the frequency of selecting the true model and SP_M of the PMSEG are larger than those of the QIC in two cases. On the other hand, the SP_C of the QIC is larger in two cases. Furthermore, by comparing Tables 1 and 2, we can see that the difference between the PMSEG and QIC is more salient when the correlation is large. We simulated several other models and obtained similar results.

Next, for the purpose of analyzing the GEE method, we consider the Mother’s Stress and Child Morbidity (MSCM) data reported in Alexander and Markowitz (1986), who studied the relationship between maternal employment and pediatric health care utilization. The MSCM data contain the information of mothers and children in the study for 28 days, and there are 167 mothers and preschool children enrolled. In this analysis, we focus on the child illness for the first 9 days. The response variable is the child illness on the study day (1=yes, 0=no), and there were eight predictor variables: Married (1=married, 0=other), Employed (1=employed, 0=unemployed), Race (child race, 1=non-white, 0=white), Household (size of household, 1=more than 3 people, 0=other), Stress (today’s mother’s stress, 1=yes, 0=no), and additionally, St1, St2 and St3 are the mother’s stress of one, two and three days before, respectively. This data have a few missing value, and we assume that the missingness mechanism is missing completely at random. We use complete 146 mothers and children data. We also assume that the response variable \( y_{ij} \) is distributed according to \( B(1, p_{ij}), \ i = 1, \ldots , 146, \ j = 1, \ldots , 9 \). For the link function, we prepare the logistic link function. For the working correlation matrix, we prepare four matrices: the working 1-dependent, autoregressive, exchangeable and independence matrices. In this analysis, we select the working correlation matrix and variables. Table 3 shows the selection probability of the model selected by minimizing the criterion and the estimated prediction error of the best model selected by the criterion. We divide the MSCM data into calibration data and validation data. The numbers of subjects in the calibration data and validation data were 136 and 10, respectively. The best subset of variables and working correlation matrices were selected by each criterion derived from the calibration data. The selection probabilities were obtained from only the calibration data. The estimated prediction errors were obtained as follows. Let \( d_t = (d_{1t}, \ldots , d_{146t})' \) be a 146-dimensional binary vector that contains 136 zeros and 10 ones at the \( t \)th iteration, \( t = 1, \ldots , 1000 \), i.e., \( d_{it} = 0 \) or 1 and \( \sum_{i=1}^{146} d_{it} = 10 \). In addition, we denote that \( \hat{\beta}_{B,[-d_t]} \) is a GEE estimator \( \hat{\beta}_{[-d_t]} \) of the best model selected from the calibration data, where \( \hat{\beta}_{[-d_t]} \) is the solution of the following equation:

\[
\sum_{i=1}^{146} (1 - d_{it}) D_i' V_i^{-1} (y_i - \mu_i) = 0_p,
\]

11
Table 3: Selection probability and estimated prediction error with the MSCM data

<table>
<thead>
<tr>
<th>Variables</th>
<th>PMSEG</th>
<th>QIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Auto</td>
<td>Ex</td>
</tr>
<tr>
<td>Race, Household, Stress, St1</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Race, Household, Stress, St3</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Race, Household, Stress, St1, St2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Race, Household, Stress, St1, St3</td>
<td>0</td>
<td>135</td>
</tr>
<tr>
<td>Race, Household, Stress, St2, St3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Married, Race, Household, Stress, St1, St3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Employed, Race, Household, Stress, St1, St3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Race, Household, Stress, St1, St2, St3</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Married, Race, Household, Stress, St1, St2, St3</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

\[
\bar{\text{PEB}}_P = 136 \times \left( \frac{1}{1000} \sum_{t=1}^{1000} \frac{1}{10} \mathcal{L}(\hat{\beta}_B,[-a_i],\hat{\beta}_f, t) - 9 \right),
\]

\[
\bar{\text{PEB}}_Q = 136 \times \frac{1}{1000} \sum_{t=1}^{1000} \frac{1}{10} \left[ -2 \sum_{i=1}^{146} \sum_{j=1}^{9} d_{it} \times Q(\hat{\beta}_B,[-a_i]; y_{ij}) \right],
\]

where \( \mathcal{L}(\beta_1, \beta_2, t) \) is defined as follows:

\[
\mathcal{L}(\beta_1, \beta_2, t) = \sum_{i=1}^{146} d_{it} \times (y_i - \mu_i(\beta_1))^\prime A_i^{-1/2}(\beta_2) R_0^{-1}(\beta_2) A_i^{-1/2}(\beta_2)(y_i - \mu_i(\beta_1)).
\]

Finally, the estimated \( \bar{\text{PEB}}_P \) and \( \bar{\text{PEB}}_Q \) are given as

\[

\bar{\text{PEB}}_P = 230.11, \quad \bar{\text{PEB}}_Q = 1018.41
\]


Finally, the estimated \( \bar{\text{PEB}}_P \) and \( \bar{\text{PEB}}_Q \) are given as

\[
\bar{\text{PEB}}_P = 230.11, \quad \bar{\text{PEB}}_Q = 1018.41
\]

Consequently, from Tables 1, 2 and 3, we recommend the use of the PMSEG rather than the QIC for selecting models in the GEE method.

5. CONCLUSION AND DISCUSSION

In the present paper, we proposed the PMSEG as a model selection criterion that reflects the correlation in the GEE method. The PMSEG is the simple criterion such as the AIC. Nowadays,
the GEE method is one of the mainstream of longitudinal analysis methods and many statistical
softwares (e.g., SAS, R, etc) support the GEE method. For these reasons, it is important to
propose a more useful criterion for analyzing longitudinal data using the GEE method.

We recall that, in deriving the PMSEG, we assume that the nuisance parameter \( \alpha \) is known.
Actually, we often estimate \( \hat{\alpha} \) because \( \alpha \) is unknown in many cases. In fact, we estimate
\( \hat{\alpha} \) in Section 4. However, Liang and Zeger (1986) showed that an estimator of \( \alpha \) is consistent under the
standard assumption, and we confirmed that the estimation of \( \alpha \) does not dramatically influence
the performance of the PMSEG from some simulation results. Theoretical study of the influence
of estimating \( \alpha \) to the PMSEG is left for the future work.

In all situations of the simulation results of Section 4, we showed that the PMSEG has better
performance than the QIC for the variable selection, and the difference between the performances
of the PMSEG and QIC is more salient when the correlation is large. Recall that the PMSEG
reflects the correlation between responses, however, the QIC is not reflective. This is probably one
of the reasons why the PMSEG has better performance than the QIC. In the study of the MSCM
data, we also showed that the PMSEG is useful as same as the QIC. Nevertheless, computational
costs of the PMSEG are lower than those of the QIC because the bias term of the PMSEG is \( 2 \times \)
the number of parameters. On the other hand, many previous studies including the QIC require
the calculation of the bias term for each candidate model. Therefore, the PMSEG is useful and
user friendly.

**APPENDIX: Derivation of (3.9)**

In this section, we calculate \( \text{Bias}_2 + \text{Bias}_4 \). \( \text{Bias}_2 \) in (3.1) can be calculated as

\[
\text{Bias}_2 = E_y \left[ \sum_{i=1}^{n} (y_i - \hat{\mu}_i)' \Sigma_{i,0}^{-1} (y_i - \hat{\mu}_i) - \sum_{i=1}^{n} (y_i - \mu_{i,0})' \Sigma_{i,0}^{-1} (y_i - \mu_{i,0}) \right] \\
= E_y \left[ 2 \sum_{i=1}^{n} (y_i - \mu_{i,0})' \Sigma_{i,0}^{-1} (\mu_{i,0} - \hat{\mu}_i) \right] + E_y \left[ \sum_{i=1}^{n} (\mu_{i,0} - \hat{\mu}_i)' \Sigma_{i,0}^{-1} (\mu_{i,0} - \hat{\mu}_i) \right]
\]

and \( \text{Bias}_4 \) in (3.1) can also be calculated as

\[
\text{Bias}_4 = E_y \left[ \sum_{i=1}^{n} (y_i - \mu_{i,0})' A_i^{-1/2} (\hat{\beta}_f) R_0^{-1} (\hat{\beta}_f) A_i^{-1/2} (\hat{\beta}_f) (y_i - \mu_{i,0}) \right] \\
- E_y \left[ \sum_{i=1}^{n} (y_i - \mu_{i,0})' A_i^{-1/2} (\hat{\beta}_f) R_0^{-1} (\hat{\beta}_f) A_i^{-1/2} (\hat{\beta}_f) (y_i - \hat{\mu}_i) \right] \\
= -E_y \left[ 2 \sum_{i=1}^{n} (y_i - \mu_{i,0})' A_i^{-1/2} (\hat{\beta}_f) R_0^{-1} (\hat{\beta}_f) A_i^{-1/2} (\hat{\beta}_f) (\mu_{i,0} - \hat{\mu}_i) \right] \\
- E_y \left[ \sum_{i=1}^{n} (\mu_{i,0} - \hat{\mu}_i)' A_i^{-1/2} (\hat{\beta}_f) R_0^{-1} (\hat{\beta}_f) A_i^{-1/2} (\hat{\beta}_f) (\mu_{i,0} - \hat{\mu}_i) \right].
\]
Thus, we obtain $\text{Bias2} + \text{Bias4}$ as follows:

\[
\text{Bias2} + \text{Bias4} = E_y \left[ 2 \sum_{i=1}^{n} (y_i - \mu_{i,0}) \{ \Sigma_{i,0}^{-1} - A_i^{-1/2}(\hat{\beta}_f) R_0^{-1}(\hat{\beta}_f) A_i^{-1/2}(\hat{\beta}_f) \} (\mu_{i,0} - \hat{\mu}_i) \right] \tag{A.1}
\]

\[
+ E_y \left[ \sum_{i=1}^{n} (\mu_{i,0} - \hat{\mu}_i) \{ \Sigma_{i,0}^{-1} - A_i^{-1/2}(\hat{\beta}_f) R_0^{-1}(\hat{\beta}_f) A_i^{-1/2}(\hat{\beta}_f) \} (\mu_{i,0} - \hat{\mu}_i) \right]. \tag{A.2}
\]

In order to calculate (A.1) and (A.2), we perform the stochastic expansion of $A_i^{-1/2}(\hat{\beta}_f)$, $R_0^{-1}(\hat{\beta}_f)$, $\mu_i(\hat{\beta}_f)$ and $\hat{\beta}_f$. Denote $D_{s,i} = A_i(\beta_s) \Delta_i(\beta_s) X_{s,i}$ and $D_{s,i,0} = A_i(\beta_s,0) \Delta_i(\beta_s,0) X_{s,i}$. Considering the same argument in section 2 and 3, $\hat{\beta}_f$ and $\mu_i(\hat{\beta}_f)$ can be expanded as

\[
\hat{\beta}_f - \beta_{s,0} = b_{f,0} + O_p(n^{-1}), \quad b_{f,0} = H_{f,n,0} s_f(n(\beta_{s,0}), \quad \mu_i(\hat{\beta}_f) - \mu_{i,0} = D_{s,i,0} b_{f,0} + O_p(n^{-1}), \tag{A.3}
\]

where $\beta_{s,0}$ is the true value of $\beta_s$, and $H_{f,n,0}$ is defined as

\[
H_{f,n,0} = \sum_{i=1}^{n} D_{s,i,0}^T A_i^{-1/2} R_0^{-1}(\alpha^*) A_i^{-1/2} D_{s,i,0}.
\]

Let $a_{f,i}(\hat{\beta}_f)$ denote the $m$-dimensional vector and the $j$th element of which is the $(j,j)$th element of $A_i^{-1/2}(\hat{\beta}_f)$. Note that $\text{diag}(a_{f,i}(\hat{\beta}_f)) = A_i^{-1/2}(\hat{\beta}_f)$. By applying a Taylor expansion around $\hat{\beta}_f = \beta_{s,0}$, $a_{f,i}(\hat{\beta}_f)$ is expanded as

\[
a_{f,i}(\hat{\beta}_f) = a_{f,i}(\beta_{s,0}) + A_{f,i,0}^* b_{f,0} + O_p(n^{-1}), \quad A_{f,i,0}^* = \frac{\partial}{\partial \beta_{s,0}} a_{f,i}(\beta_{s,0}) \bigg|_{\beta_{s} = \beta_{s,0}}.
\]

Hence, we obtain

\[
A_i^{-1/2}(\hat{\beta}_f) = \text{diag}(a_{f,i}(\hat{\beta}_f)) = A_i^{-1/2} + \text{diag}(A_{f,i,0}^* b_{f,0}) + O_p(n^{-1}). \tag{A.4}
\]

Note that $b_{f,0}, D_{s,i,0} b_{f,0}, \text{diag}(A_{f,i,0}^* b_{f,0}) = O_p(n^{-1/2})$. Moreover, substituting (A.3) and (A.4) into $R_0(\hat{\beta}_f)$ yields following:

\[
R_0(\hat{\beta}_f) = -\frac{1}{n} \sum_{i=1}^{n} A_{i,0}^{-1/2} \{ D_{s,i,0} b_{f,0} (y_i - \mu_{i,0})' + (y_i - \mu_{i,0}) (D_{s,i,0} b_{f,0})' \} A_{i,0}^{-1/2}
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} A_{i,0}^{-1/2} (y_i - \mu_{i,0})(y_i - \mu_{i,0})' A_{i,0}^{-1/2}
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} \text{diag}(A_{f,i,0}^* b_{f,0})(y_i - \mu_{i,0})(y_i - \mu_{i,0})' A_{i,0}^{-1/2}
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} A_{i,0}^{-1/2} (y_i - \mu_{i,0})(y_i - \mu_{i,0})' \text{diag}(A_{f,i,0}^* b_{f,0}) + O_p(n^{-1}). \tag{A.5}
\]
By the Lindeberg central limit theorem, the first term of (A.5) is $O_p(n^{-1})$. Thus, using this fact and (A.5), we obtain

$$R_0^{-1/2} R_0 (\hat{\beta}_f) R_0^{-1/2} = I_m - R_0^{-1/2} \left\{ R_0 - \frac{1}{n} \sum_{i=1}^{n} A_{i,0}^{-1/2} (y_i - \mu_{i,0})(y_i - \mu_{i,0})' A_{i,0}^{-1/2} 
- \frac{1}{n} \sum_{i=1}^{n} \text{diag}(A_{i,0}^* b_{f,0})(y_i - \mu_{i,0})(y_i - \mu_{i,0})' A_{i,0}^{-1/2} 
- \frac{1}{n} \sum_{i=1}^{n} A_{i,0}^{-1/2} (y_i - \mu_{i,0})(y_i - \mu_{i,0})' \text{diag}(A_{i,0}^* b_{f,0}) \right\} R_0^{-1/2} + O_p(n^{-1}).$$

Therefore, by calculating the inverse matrix of $R_0^{-1/2} R_0 (\hat{\beta}_f) R_0^{-1/2}$, $R_0^{-1}(\hat{\beta}_f)$ can be expanded as

$$R_0^{-1}(\hat{\beta}_f) = R_0^{-1} + R_0^{-1} \left\{ R_0 - \frac{1}{n} \sum_{i=1}^{n} A_{i,0}^{-1/2} (y_i - \mu_{i,0})(y_i - \mu_{i,0})' A_{i,0}^{-1/2} 
- \frac{1}{n} \sum_{i=1}^{n} \text{diag}(A_{i,0}^* b_{f,0})(y_i - \mu_{i,0})(y_i - \mu_{i,0})' A_{i,0}^{-1/2} 
- \frac{1}{n} \sum_{i=1}^{n} A_{i,0}^{-1/2} (y_i - \mu_{i,0})(y_i - \mu_{i,0})' \text{diag}(A_{i,0}^* b_{f,0}) \right\} R_0^{-1} + O_p(n^{-1}).$$

(A.6)

Note that the second term of (A.6) is $O_p(n^{-1/2})$.

Next, we calculate (A.2). By using (A.4) and (A.6), we obtain

$$\Sigma_{i,0}^{-1} - A_{i}^{-1/2}(\hat{\beta}_f) R_0^{-1}(\hat{\beta}_f) A_{i}^{-1/2}(\hat{\beta}_f)
= - \text{diag}(A_{i,0}^* b_{f,0}) R_0^{-1} A_{i,0}^{-1/2} - A_{i,0}^{-1/2} R_0^{-1} \text{diag}(A_{i,0}^* b_{f,0}) 
- A_{i,0}^{-1/2} R_0^{-1} \left\{ R_0 - \frac{1}{n} \sum_{i=1}^{n} A_{i,0}^{-1/2} (y_i - \mu_{i,0})(y_i - \mu_{i,0})' A_{i,0}^{-1/2} 
- \frac{1}{n} \sum_{i=1}^{n} \text{diag}(A_{i,0}^* b_{f,0})(y_i - \mu_{i,0})(y_i - \mu_{i,0})' A_{i,0}^{-1/2} 
- \frac{1}{n} \sum_{i=1}^{n} A_{i,0}^{-1/2} (y_i - \mu_{i,0})(y_i - \mu_{i,0})' \text{diag}(A_{i,0}^* b_{f,0}) \right\} R_0^{-1} A_{i,0}^{-1/2} + O_p(n^{-1}).$$

(A.7)

Note that $\Sigma_{i,0}^{-1} - A_{i}^{-1/2}(\hat{\beta}_f) R_0^{-1}(\hat{\beta}_f) A_{i}^{-1/2}(\hat{\beta}_f) = O_p(n^{-1/2})$ and $D_{i,0} b_{1,0} = O_p(n^{-1/2})$. Therefore, by substituting (3.4) and (A.7) into (A.2), we obtain

$$\text{(A.2)} = O(n^{-1}).$$

(A.8)

Recall that, in general, a moment of statistic can be expanded as power series in $n^{-1}$. Hence, the order of (A.8) is shown by $O(n^{-1})$, not $O(n^{-1/2})$. 

15
Finally, we calculate (A.1). By substituting (3.4) and (A.7) into (A.1), we obtain

$$
(A.1) = E_y \left[ 2 \sum_{i=1}^{n} \kappa'_{i,0} (\text{diag}(A^{*}_{f,i,0} b_{f,i,0}) R^{-1}_{0} A^{-1/2}_{i,0} + A^{-1/2}_{i,0} R^{-1}_{0} \text{diag}(A^{*}_{f,i,0} b_{f,i,0})) D_{i,0} b_{1,0} \right]
- E_y \left[ 2 \sum_{i=1}^{n} \kappa'_{i,0} A^{-1/2}_{i,0} R^{-1}_{0} \frac{1}{n} \sum_{j=1}^{n} A^{-1/2}_{j,0} \kappa_{j,0} A^{-1/2}_{j,0} R^{-1}_{0} A^{-1/2}_{i,0} D_{i,0} b_{1,0} \right]
- E_y \left[ 2 \sum_{i,j}^{n} \kappa'_{i,0} A^{-1/2}_{i,0} R^{-1}_{0} \text{diag}(A^{*}_{f,j,0} b_{f,j,0}) \kappa_{j,0} A^{-1/2}_{j,0} R^{-1}_{0} A^{-1/2}_{i,0} D_{i,0} b_{1,0} \right]
- E_y \left[ 2 \sum_{i,j}^{n} \kappa'_{i,0} A^{-1/2}_{i,0} R^{-1}_{0} \kappa_{j,0} \kappa_{j,0} A^{-1/2}_{j,0} \text{diag}(A^{*}_{f,j,0} b_{f,j,0}) R^{-1}_{0} A^{-1/2}_{i,0} D_{i,0} b_{1,0} \right]
+ E_y \left[ 2 \sum_{i=1}^{n} \kappa'_{i,0} \sum_{i,0}^{1} D_{i,0} b_{i,0} \right] + O(n^{-1}).
$$

(A.9)

Here, \( \kappa_{i,0} = (y_i - \mu_{i,0}) \) and \( \sum_{i,j} = \sum_{i=1}^{n} \sum_{j=1}^{n} \). Moreover, in order to simplify the calculation, we define the following notation:

$$
\sum_{i\neq j} = \sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n}, \quad b_{f,i,0} = H^{-1}_{j,n,0} D^*_{i,j,0} A^{-1}_{i,0} \kappa_{i,0}.
$$

Recall that \( E[\kappa_{i,0} \otimes \kappa'_{j,0} \kappa_{k,0}] = 0 \), (not \( i = j = k \)). Hence, the first term of (A.9) is as follows:

$$
E_y \left[ 2 \sum_{i=1}^{n} \kappa'_{i,0} \text{diag}(A^{*}_{f,i,0} b_{f,i,0}) R^{-1}_{0} A^{-1/2}_{i,0} + A^{-1/2}_{i,0} R^{-1}_{0} \text{diag}(A^{*}_{f,i,0} b_{f,i,0}) \right] D_{i,0} b_{1,0} \right]
= E_y \left[ 2 \sum_{i=1}^{n} \kappa'_{i,0} \text{diag}(A^{*}_{f,i,0} b_{f,i,0}) R^{-1}_{0} A^{-1/2}_{i,0} + A^{-1/2}_{i,0} R^{-1}_{0} \text{diag}(A^{*}_{f,i,0} b_{f,i,0}) \right] D_{i,0} b_{1,0} \right]
= O(n^{-1}).
$$

(A.10)

Similarly, since \( E_y[\kappa'_{i,0} \kappa_{j,0} \kappa'_{j,0} \kappa_{k,0}] = 0 \) unless \( i = k \), the second term of (A.9) can be calculated as

$$
- E_y \left[ 2 \sum_{i=1}^{n} \kappa'_{i,0} A^{-1/2}_{i,0} R^{-1}_{0} \frac{1}{n} \sum_{j=1}^{n} A^{-1/2}_{j,0} \kappa_{j,0} A^{-1/2}_{j,0} R^{-1}_{0} A^{-1/2}_{i,0} D_{i,0} b_{1,0} \right]
- E_y \left[ 2 \sum_{i=1}^{n} \kappa'_{i,0} A^{-1/2}_{i,0} R^{-1}_{0} \frac{1}{n} \sum_{j=1, i \neq j}^{n} A^{-1/2}_{j,0} \kappa_{j,0} A^{-1/2}_{j,0} R^{-1}_{0} A^{-1/2}_{i,0} D_{i,0} b_{1,0} \right] + O(n^{-1})
= - E_y \left[ 2 \sum_{i=1}^{n} \kappa'_{i,0} \sum_{i,0}^{1} D_{i,0} b_{i,0} \right] + O(n^{-1}) = -2p + O(n^{-1}).
$$

(A.11)

Note that \( -2p \) in (A.11) is obtained from (3.6). Furthermore, \( E_y[\kappa'_{i,0} (\kappa_{j,0} \otimes \kappa'_{k,0})(\kappa_{k,0} \otimes \kappa_{l,0})] = 0 \) expect in the following cases:

\( i = j = l \) or \( i = j \neq k = l \) or \( i = l \neq k = j \) or \( j = l \neq k = i \).
Thus, the third term of (A.9) is given by

\[
-E_y \left[ \frac{2}{n} \sum_{i,j} \kappa_{i,0}^' A_{i,0}^{-1/2} R_0^{-1} \text{diag}(A_{f,j,0}^* b_{f,0}) \kappa_{i,0}^' A_{j,0}^{-1/2} R_0^{-1} A_{i,0}^{-1/2} D_{i,0} b_{1,0} \right]
\]

\[
= -E_y \left[ \frac{2}{n} \sum_{i=1}^n \kappa_{i,0}^' A_{i,0}^{-1/2} R_0^{-1} \text{diag}(A_{f,i,0}^* b_{f,0}) \kappa_{i,0}^' A_{i,0}^{-1/2} R_0^{-1} A_{i,0}^{-1/2} D_{i,0} b_{1,0} \right]
\]

\[
- E_y \left[ \frac{2}{n} \sum_{i \neq j} \kappa_{i,0}^' A_{i,0}^{-1/2} R_0^{-1} \text{diag}(A_{f,j,0}^* b_{f,0}) \kappa_{j,0}^' A_{j,0}^{-1/2} R_0^{-1} A_{i,0}^{-1/2} D_{i,0} b_{1,0} \right] + O(n^{-1})
\]

\[
= O(n^{-1}). \tag{A.12}
\]

In the same manner as in the calculation of the third term of (A.9), the fourth term of (A.9) is calculated as

\[
-E_y \left[ \frac{2}{n} \sum_{i,j} \kappa_{i,0}^' A_{i,0}^{-1/2} R_0^{-1} \text{diag}(A_{f,j,0}^* b_{f,0}) \kappa_{i,0}^' A_{j,0}^{-1/2} R_0^{-1} A_{i,0}^{-1/2} D_{i,0} b_{1,0} \right] = O(n^{-1}). \tag{A.13}
\]

Moreover, the fifth term of (A.9) is obtained from (3.6), as follows:

\[
E_y \left[ 2 \sum_{i=1}^n \kappa_{i,0}^' \Sigma_{i,0}^{-1} D_{i,0} b_{1,0} \right] = 2p. \tag{A.14}
\]

Substituting (A.10), (A.11), (A.12), (A.13) and (A.14) into (A.9), (A.1) is given by

\[
(A.1) = O(n^{-1}). \tag{A.15}
\]

Consequently, from (A.8) and (A.15), we obtain \(\text{Bias}^2 + \text{Bias}^4 = O(n^{-1})\).

**ACKNOWLEDGEMENTS**

The authors would like to thank Dr. Hirofumi Wakaki of Hiroshima University for their helpful comments and suggestions.

**REFERENCES**


